Analysis of Airline Delay and Cancellation Data, 2009 - 2018

Paul Smith

Western Governors University

**Table of Contents**

[A. Project Overview 3](#_Toc128098450)

[A1. Research Question or Organizational Need 3](#_Toc128098451)

[A2. Context and Background 3](#_Toc128098452)

[A3. Summary of Published Works 3](#_Toc128098453)

[A3a. Relation of Published Works to Project 4](#_Toc128098454)

[A4. Summary of Data Analytics Solution 4](#_Toc128098455)

[A5. Benefit to Organization and Decision-Making Process 5](#_Toc128098456)

[B. Data Analytics Plan 5](#_Toc128098457)

[B1. Goals, Objectives, and Deliverables 5](#_Toc128098458)

[B2. Scope of Project 6](#_Toc128098459)

[B3. Standard Methodology 6](#_Toc128098460)

[B4. Timeline and Milestones 7](#_Toc128098461)

[B5. Resources and Costs 7](#_Toc128098462)

[B6. Criteria for Success 8](#_Toc128098463)

[C. Design of Data Analytics Solution 8](#_Toc128098464)

[C1. Hypothesis 8](#_Toc128098465)

[C2. Analytical Method 9](#_Toc128098466)

[C2a. Justification of Analytical Method 9](#_Toc128098467)

[C3. Tools and Environments of Solution 9](#_Toc128098468)

[C4. Methods and Metrics to Evaluate Statistical Significance 9](#_Toc128098469)

[C4a. Justification of Methods and Metrics 9](#_Toc128098470)

[C5. Practical Significance 10](#_Toc128098471)

[C6. Visual Communication 10](#_Toc128098472)

[D. Description of Datasets 10](#_Toc128098473)

[D1. Source of Data 10](#_Toc128098474)

[D2. Appropriateness of Dataset 11](#_Toc128098475)

[D3. Data Collection Methods 11](#_Toc128098476)

[D4. Data Quality 11](#_Toc128098477)

[D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance 12](#_Toc128098478)

[D5a. Precautions 12](#_Toc128098479)

[E. Sources 13](#_Toc128098480)

# Project Overview

# A1. Research Question or Organizational Need

This project will analyze airline delay and cancellation data from 2009 to 2018 in order to determine the whether the overall cause of flight delays is late aircraft delay. Flight delays cost the US economy a great deal of money - $32.9 billion in 2007 (Ball, M. et al. 2010), and determining the cause of these delays could help to substantially reduce this cost.

# A2. Context and Background

Ball, M. et al. (2010), sponsored by the Federal Aviation Administration (FAA), analyzed a variety of costs caused by flight delays. This included the cost to airlines, cost to passengers, cost of lost demand, as well as the indirect impact of delay on the US economy. The report concluded that the total cost of all US air transportation delays in 2007 was $32.9 billion. Clearly, flight delays are a serious and widespread problem in the US.

An exploration of multi-year (2009 – 2018) airline delay and cancellation data will be made to determine whether the overall cause of airline delays is late aircraft delay.

# A3. Summary of Published Works

Weather, especially extreme weather, has an impact on flights. Goodman and Small Griswold (2019) described the average impact that various weather phenomena have on aircraft delays. Assessing airports individually is important, as differences in weather and airline efficiencies impact their operations efficiencies. Differences in airport weather climates need to be understood, so that the impact of inclement weather on efficiency can be determined. Finally, the impact of weather on delays and cancellation needs to be reconciled with climatological weather patterns.

Rupp (2007) suggests that flight delays should be investigated from both the airline and passenger perspectives. Often, flight delay investigations focus on the airline perspective. Defining the lateness of aircraft as only happening when an aircraft arrives more than fifteen minutes after its scheduled landing time is one of the ways in which airlines ignore the passenger perspective. The airline looks at excess travel time, whereas the passengers see arrival or departure delays. Significant effects on flight delays include seating capacity, load factor, departure time, and distance. Once airport-specific effects are controlled, most estimations show that airport concentration at origination have longer departure and arrival delays.

Blackwood (2012) wrote that flight delays negatively impact the environment, the economy, and society. He concluded after studying various delays at Chicago O’Hare airport that late aircraft accounted for forty percent of all delays, carrier related issues accounted for twenty-nine percent of all delays, national airspace bottlenecks accounted for twenty-five percent of all delays, weather accounted for six percent of all delays, and finally, security issues accounted for one percent of all delays.

# A3a. Relation of Published Works to Project

It is possible that weather is one of the causes of flight delays, or even the largest cause of flight delays. This project must look at delays and determine the effect weather has on delays.

Passengers are more affected by flight delays than any other group, so this investigation will focus on the passenger perspective for defining flight delays, rather than the airline perspective.

It will be interesting to see if the causes of delays at a single large airport are indicative of delays across the entire country.

# A4. Summary of Data Analytics Solution

This project will analyze airline delay and cancellation data from 2009 to 2018 and determine the causes of flight delays. A Jupyter Notebook will be created that will contain Python code (and markdown text to explain the code and results) within its cells to load the data, cleanse the data, perform the analysis and conclude whether the null hypothesis was proven or disproven.

The data will be loaded from CSV data files downloaded from Kaggle containing one year of data per file. This data will be concatenated into a single dataset.

This single dataset will be cleansed. The pattern of NaNs (Not a Number values that represent missing data or sometimes zero depending on their context) will be determined. Any columns containing only NaNs will be removed from the dataset. Cancelled and diverted flights will be removed from the dataset since these are not delayed flights. NaNs are replaced with zeroes in the five delay columns since these represent a delay of zero minutes. Rows that contain NaNs in columns showing less than two percent of NaNs are removed as these represent missing data. The dataset is made more readable by replacing the carrier codes with actual airline business names. Non-delayed flights are removed from the dataset, since this project is concerned with delayed flights, and finally the flight date column is converted to a DateTime type to allow for future analysis.

The data analysis is performed. A correlation matrix heatmap is generated and multicollinear columns are removed from the dataset. The correlation matrix heatmap is redisplayed, and correlations are noted. Each correlation is tested for statistical significance and a plot is generated of the number of delays that correspond to the columns with statistically significant correlations versus the flight dates. This plot is inspected, and the causes of delays are ranked.

A conclusion is determined as to whether the null hypothesis was proven or disproven.

# A5. Benefit to Organization and Decision-Making Process

The benefits of this analysis will help both the airline decision makers and passengers. These groups are the stakeholders for this project. For airline decision makers, the cause of delays will be identified, and possible mitigation strategies can be formulated that will likely reduce costs and generate passenger goodwill. For passengers, knowing the cause of delays will allow them to pressure the airlines to execute mitigation strategies, e.g., buying tickets for airlines that have reduced flight delays compared to their peers.

# Data Analytics Plan

# B1. Goals, Objectives, and Deliverables

The goal of this project is to find the cause of flight delays. To do this a Jupyter Notebook will be used to perform data analysis on the flight delay and cancellation data from 2009 – 2018.

The objectives for this goal are:

* Concatenate the data into a single dataset, so that the data analysis can be performed on a single dataset.
  + The deliverable is to return a single dataset containing all the years of data.
* Cleanse the dataset, so that missing or unknown data does not compromise the results.
  + The deliverable is to return a single dataset free from unknown or missing data.
* Analyze the dataset for the cause of flight delays.
  + The deliverable is to list the cause of flight delays.

# B2. Scope of Project

The scope of this project will include a Jupyter Notebook. This notebook will provide various steps in order to analyze the flight delay and cancellation dataset. The output of this notebook will be the results of each step culminating in the cause of flight delays. The scope of this project will not include analyzing cancelled or diverted flights, nor will it involve breaking down any flight delay causes further, e.g., if weather is a cause of flight delays, then the data will not be analyzed to discover which airports are more affected by bad weather conditions.

# B3. Standard Methodology

The Waterfall project methodology will be used by this project. It emphasizes that a project progresses from the beginning to the end of a project in a linear progression. The phases of this project are Requirements, Design, Implementation, Verification, and Maintenance. The following explains how the project will proceed during each of these phases.

**Requirements:** All customer requirements are gathered before any other phase is begun. In this phase, the project scope is determined, the user expectations are decided, and the resources needed to complete the project are finalized.

**Design:** The tasks needing to be completed, in order to achieve the project objectives, are determined in this phase. Some of these tasks include determining what data cleansing will be necessary, the steps needed to analyze the dataset, and the visualizations required for the results.

**Implementation:** The tasks needed to achieve the objectives and test the Jupyter notebook to ensure it is producing the desired results are completed in this phase.

**Verification:** In this stage, I will complete a standalone file for this project, so that it can be implemented by anyone else who has access to suitable hardware and software, i.e., a Jupyter development environment is installed.

**Maintenance:** This stage will not apply to this project, as it will not be in production in any companies. However, it could be uploaded to Kaggle, and in that case bug fixes and modifications could be requested.

# B4. Timeline and Milestones

Present a table showing for each milestone its projected start and end dates, and its projected duration:

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **Projected Start Date** | **Projected End Date** | **Duration (hours)** |
| Establish requirements for analytics process | 03/01/2024 | 03/03/2024 | 24 |
| Download dataset | 03/04/2024 | 03/04/2024 | 2 |
| Code notebook – loading data | 03/04/2024 | 03/04/2024 | 6 |
| Code notebook – cleansing data | 03/05/2024 | 03/08/2024 | 32 |
| Code notebook – data analysis | 03/09/2024 | 03/14/2024 | 48 |
| Test notebook | 03/15/2024 | 03/18/2024 | 32 |
| Create html file showing all notebook code and results | 03/19/2024 | 03/19/2024 | 1 |

# B5. Resources and Costs

|  |  |
| --- | --- |
| **Personnel, technology, or infrastructure** | **Cost** |
| Dataset for data analysis | N/A |
| Jupyter Notebook development environment | N/A |
| 145 hours | N/A |

The resources needed for this project to be completed and implemented are limited to a dataset for analysis (downloaded for free from Kaggle), my time (145 hours of conception, design, coding, and testing), my computer for development (no cost, it has already been acquired). There are no additional resources or costs that will be associated with this project.

# B6. Criteria for Success

The criteria I will use for success are that the causes of flight delays are identified, any identified causes are statistically significant, and that those causes are ranked in terms of greatest number of flight delays to smaller number of flight delays.

|  |  |  |
| --- | --- | --- |
| **Criterion/Metric** | **Required Data** | **Cut Score for Success** |
| Are causes of flight delays identified? | Correlation Matrix heatmap | Success is if and only if there is a correlation for each cause |
| Are causes statistically significant? | T-test output | Success is if and only if there is a statistically significance for each cause |
| Are the number of delays for each cause ranked? | Plot of number of delays versus flight date | Success if and only if each cause is ranked relative to the others |

# Design of Data Analytics Solution

# C1. Hypothesis

The null hypothesis, H0, for this analysis is ‘late aircraft delay is not the biggest cause of flight delays.’

# C2. Analytical Method

The analytical method I will use in my data analysis will be descriptive. The data analysis technique I will use is multivariate analysis in the form of a correlation matrix. Each result will undergo a t-test to ensure its statistical significance. The significant results will be graphed to show number of delays versus flight date to allow a ranking of flight delay causes. The ranking of late aircraft delay will be determined.

# C2a. Justification of Analytical Method

I chose this method and technique because there are several possible variables that may depend on each other. The correlation matrix enables those variables that are correlated to be identified easily. The t-test allows the statistically significant results to be determined, and the graph of number of delays versus flight date enables the causes of flight delays to be ranked, so that the ranking of late aircraft delay is determined.

# C3. Tools and Environments of Solution

A Jupyter Notebook will be used to manipulate the data, because it provides a robust environment to manipulate and analyze the dataset. This tool also provides a flexible way to output results, as both text and graphical plots can be generated. Also, the notebook can be distributed reasonably easy, so that other researchers can perform the same analysis on similar datasets.

# C4. Methods and Metrics to Evaluate Statistical Significance

Statistical significance for the correlation coefficients of flight delay causes will be shown via a t-test. A null hypothesis will be proposed for each flight delay cause, and an attempt will be made to disprove that hypothesis, thereby proving that correlation coefficient is statistically significant. P-value of less than 0.05 will show that the result is significant.

# C4a. Justification of Methods and Metrics

This approach is the most appropriate because it will allow me to demonstrate whether a flight delay cause is statistically significant.

# C5. Practical Significance

The practical significance of knowing what causes flight delays allows mitigation strategies to be formulated, and whether attempting mitigation is useful. For example, if weather was the biggest cause of flight delays, then it would be impractical to try and mitigate the effect of weather on delays. It would then be more productive to try and mitigate other causes of delays. Delays that are under an airline’s control could be reduced to provide better customer service and goodwill. For example, allowing planes to fly faster when late would help reduce or negate late arrivals. This is something that is within the power of airline decision makers. Knowing the ranking of flight delay causes allows decision makers to determine the best use of resources in reducing flight delays.

# C6. Visual Communication

A correlation matrix heatmap will be used to communicate the findings, as well as a plot of the number of delays versus the date of each flight. The heatmap is very good at displaying many correlation coefficients compared to other methods in a fixed space. The plot is a very good method of comparing the number of delays for each delay type in a fixed space so that their relative ranking is easy to determine, particularly for large datasets.

# Description of Datasets

# D1. Source of Data

Airline Delay and Cancellation Data 2009 - 2018 were collected from data provided by Kaggle. This website provides various publicly downloadable datasets. This data has been combined from multiple US Government (Bureau of Transportation Statistics) datasets.

There are ten data files associated with this project:

2009.csv: Airline Delay and Cancellation Data dataset for 2009 (Milan, 2020).

2010.csv: Airline Delay and Cancellation Data dataset for 2010 (Milan, 2020).

2011.csv: Airline Delay and Cancellation Data dataset for 2011 (Milan, 2020).

2012.csv: Airline Delay and Cancellation Data dataset for 2012 (Milan, 2020).

2013.csv: Airline Delay and Cancellation Data dataset for 2013 (Milan, 2020).

2014.csv: Airline Delay and Cancellation Data dataset for 2014 (Milan, 2020).

2015.csv: Airline Delay and Cancellation Data dataset for 2015 (Milan, 2020).

2016.csv: Airline Delay and Cancellation Data dataset for 2016 (Milan, 2020).

2017.csv: Airline Delay and Cancellation Data dataset for 2017 (Milan, 2020).

2018.csv: Airline Delay and Cancellation Data dataset for 2018 (Milan, 2020).

# D2. Appropriateness of Dataset

These data are appropriate for addressing the project goal since they contain the proper fields for analysis and are specifically designed for flight analysis. These data are the most appropriate way to provide my analysis with the data as they are freely provided by the US Government and created by a department whose expertise lies in transport. This means that I will not need to use any propriety data and deal with any legal ramifications that my entail.

# D3. Data Collection Methods

The data was collected by downloading it from Kaggle. The advantage of collecting data this way is that some data cleansing had taken place, since this data ultimately originated from US Government sources. The data quality is good too. The disadvantage of collecting data this way is that some of the data was not needed, e.g., cancelled flights, so further cleansing was required. The data was examined and cleansed to ensure that it was as accurate as possible and of high quality, i.e., no missing data was used to contaminate the results. An account with Kaggle had to be created before the data could be downloaded.

# D4. Data Quality

The data had some quality issues. One column of data was unusable. It had no useful column name and contained no actual data. This column was removed from the dataset. Some data that required a zero value did not contain that value and that had to be corrected. Some data was assigned the wrong type when loaded. This was corrected. There were several rows that contained unknown or missing data. These rows were removed from the dataset. For the purposes of this analysis all data relating to cancelled flights and diverted flights were removed.

# D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance

The data that were used are publicly available to anyone, and do not contain any personally identifiable information. While the data should be under the control of internal standards for data governance, there are few issues that would be caused if the data was stolen, since it is publicly available.

# D5a. Precautions

When the dataset was modified, tests were performed to confirm that any modifications were correctly executed. During analysis, each step was checked to make sure that the results obtained for each step were correct. The project files (data files, Jupyter Notebook, PDF files) were stored on local storage, but at the end of each day were pushed to GitHub in case of a local issue. The project files are locally stored under a user account that is password protected. Dissemination of the Jupyter Notebook would be performed under a suitable license, e.g., the MIT license would allow dissemination of the notebook to almost anyone.

# Sources

Ball, M. and Barnhart, C. and Dresner, M. and Hansen, M. and Neels, K. and Odoni, A. and Peterson, E. and Sherry, L. and Trani, A. and Zou, B. (2010, October 16). *Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States.* Institute of Transportation Studies, University of California, Berkeley. https://worldcat.org/title/671248487

Rupp, N. G. (2007, May 20). *Further Investigations into the Causes of Flight Delays*. East Carolina University. <https://economics.ecu.edu/wp-content/pv-uploads/sites/165/2019/07/ecu0707.pdf>

Blackwood, P. (2012, April). *Understanding Flight Delays at U.S. Airports in 2010, Using Chicago O’Hare International Airport as a Case Study*. Western Michigan University <https://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=1033&context=masters_theses>

Goodman, C. J. and Small Griswold, J. D. (2019, January 4). *Meteorological Impacts on Commercial Aviation Delays and Cancellations in the Continental United States*. Journal of Applied Meteorology and Climatology. <https://journals.ametsoc.org/view/journals/apme/58/3/jamc-d-17-0277.1.xml?tab_body=pdf>

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2009.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2010.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2011.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2012.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*.Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2013.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2014.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2015.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2016.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2017.csv

Milan, T. (2020). *Airline Delay and Cancellation Data 2018*. Kaggle. https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2018.csv